BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 9: e182910
doi: 10.3897/biss.9.182910

OPEN ACCESS

Conference Abstract

LLM-Based Transcription of Botanical Catalogues for Digital Access to Special Collections

Ishaipiriyan Karunakularatnam‡, Nicky Nicolson‡, Priscila Reis‡, Magda Upton‡

‡ Royal Botanic Gardens, Kew, London, United Kingdom

Corresponding author: Ishaipiriyan Karunakularatnam (i.karunakularatnam2@kew.org), Nicky Nicolson (n.nicolson@kew.org), Priscila Reis (p.reis@kew.org), Magda Upton (m.steele@kew.org)

Received: 18 Dec 2025 | Published: 23 Dec 2025

Citation: Karunakularatnam I, Nicolson N, Reis P, Upton M (2025) LLM-Based Transcription of Botanical Catalogues for Digital Access to Special Collections. Biodiversity Information Science and Standards 9: e182910. https://doi.org/10.3897/biss.9.182910

Abstract

The Kew Herbarium houses over seven million specimens and represents one of the world's most significant botanical collections. Recent large-scale digitisation has made over 6.4 million specimens (Royal Botanic Gardens, Kew 2025) globally accessible, supporting active research into biodiversity and ecology. However, this progress has primarily been focused on herbarium sheets, leaving many specialist collections under-digitised and difficult to access. These specialist collections, stored in folders of specimen sheets, are often accompanied by printed catalogues containing the key metadata. Digitisation of these collections is difficult to process using standard herbarium workflows (which are scaled efficiently across millions of specimen sheets) due to their highly varying structure and format. This diversity makes automation more challenging so transcription is often manual, which is slow, expensive, and prone to errors. As main herbarium digitisation nears completion, adaptable, cost-effective methods are needed for specialist catalogues.

This project has developed a modular, semi-automated pipeline to extract structured data from scanned pages of botanical catalogues (images/PDFs) using the Lightfoot catalogue (Bowden 1989) for development and testing. This pipeline allows user configuration across different sub-collections and extraction models. It uses Optical Character Recognition (OCR) and Large Language Models (LLMs) based transcription, allowing adaptation to different catalogue formats through prompt changes,

© Karunakularatnam I et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

without the need for costly custom pipelines. It consists of four stages: image processing; text extraction; data structuring; and transcription.

Provided with the scanned pages, image processing (with DeepDoctection (Meyer 2024)) is performed to identify and crop regions of interest (ROI) which are normalised for text extraction. Text extraction is performed on each ROI, and the extracted text is saved sequentially. Tesseract OCR (Smith 2007) is used by default due to its adaptability and strong community support, but support for LLMs with visual reasoning, such as Qwen2.5 (Yang et al. 2024), is also implemented. Before transcription, the extracted data is cleaned and chunked with respect to family names and species names with the help of TaxoNERD (Le Guillarme and Thuiller 2022). During transcription, species-level text chunks are passed to a LLM with user-defined instructions (based on catalogue) and the pre-defined structure of output. The output from this previous step is saved in both JSON and CSV formats.

Evaluation on the Lightfoot catalogue yielded F1-scores of 86%, 87% and 91% for family names, species names and folder descriptions, respectively, indicating strong extraction performance. Additionally folder descriptions had Word Error Rate (WER) and Character Error Rate (CER) of 24% and 18% reflecting reasonable baseline accuracy given the variable quality of the text.

Despite this, key challenges remain: OCR accuracy drops on degraded or low-resolution scans, and LLM hallucinations can misclassify species names or description fields. These issues highlight the need for a quality control layer, such as curator validation, and refining prompt strategies to reduce errors. Nevertheless, this pipeline demonstrates that significant efficiency gains can be achieved while reducing reliance on manual data entry.

This project provides an operational workflow for the digitisation of herbarium catalogue data using OCR and LLMs. Future work will focus on improving OCR accuracy, refining LLM prompts and chunking strategy. Additional testing will be performed on the Hanbury and other specialist catalogues to further validate the pipeline's performance on varying formats. Ultimately, this approach contributes to the broader goal of accelerating digitisation across herbaria, ensuring valuable historical records are preserved and made accessible for future research.

Keywords

herbarium digitisation, specialist catalogues, OCR, Tesseract, biodiversity informatics

Presenting author

Ishaipiriyan Karunakularatnam

Presented at

Living Data 2025

Acknowledgements

Many thanks to Marie-Helene Weech and Ashleigh Whittaker for their valuable guidance and helpful discussions.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Bowden JK (1989) John Lightfoot, His Work and Travels: With a Biographical Introduction and a Catalogue of the Lightfoot Herbarium. Royal Botanic Gardens Kew
- Le Guillarme N, Thuiller W (2022) TaxoNERD: Deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. Zenodo https://doi.org/10.5281/zenodo.13533134
- Meyer J, et al. (2024) DeepDoctection. v0.46.2. URL: https://github.com/deepdoctection/deepdoctection
- Royal Botanic Gardens, Kew (2025) Royal Botanic Gardens, Kew - Herbarium Specimens. Occurrence dataset https://doi.org/10.15468/ly60bx
- Smith R (2007) An Overview of the Tesseract OCR Engine. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2629-633. https://doi.org/10.1109/icdar.2007.4376991
- Yang A, Li A, Yang B, Zhang B, et al. (2024) Qwen2 Technical Report. arXiv. arXiv: 2407.10671 [cs]. https://doi.org/10.48550/arXiv.2407.10671